Lateral

Journal of the Cultural Studies Association

Call the (Bot-)Police: User-Led Platform Governance of "(In)Authenticity" on Instagram

by Nathalie Schäfer | Digital Platforms and Agency, Issue 14.2 (Fall 2025)

ABSTRACT The pervasive practice of botting by using fame-enhancing bots and operating porn bots, and Instagram's opaque and unreliable authenticity governance, has evoked human Instagram users to actively police and govern botting and other bot activity. Botting describes repetitive and quantitative posting, messaging and engaging on social media platforms to provoke reciprocal engagement. Running bot police accounts, Instagram users engage and try to play an active role in the authenticity governance process of Instagram. The article investigates why and how Instagram users govern two types of "inauthentic" Instabots, which concentrate on detecting automated interactions by fame-enhancing bots and reporting porn bots. Two case studies of three Instagram accounts and their content show different approaches to user-led authenticity governance and how these profiles perceive themselves as custodians of "authenticity." The key findings reveal that their activities of pillorying and collective flagging are considered digital vigilantism that potentially cause harm to other users. Furthermore, they show mechanisms of peer surveillance and mutual moderation on a user level, and complex power asymmetries between the platform company, its moderation systems, advertisers, and its users.

KEYWORDS Instagram, authenticity, botting, governance, surveillance

Introduction

"Too much harmless content gets censored, too many people find themselves wrongly locked up in 'Facebook jail,'" said Joel Kaplan, Meta's Chief Global Affairs Officer, on January 7, 2025, to justify the end of independent third-party fact-

checking at Meta.[1] This statement highlights the core problem of Meta's platform governance. Surprisingly, the company admits to the failure of its content moderation and platform governance. However, Meta neglects the other side of the issue: too little content containing fake news, hate speech, or porn bots is censored or "locked up in Facebook jail." This further illustrates their unwillingness to act as custodians of their platforms and accept the responsibility of ensuring a safe online environment. The same applies to Instagram, Facebook's associated platform under Meta. By introducing community notes, the company is shifting responsibility to their users despite the existence of user-level governance features, such as flagging, that have long been established on the platform. Meta has allowed the development of practices for user-led policing of "inauthentic" behavior. The goal was to lock away fame-enhancing bots and porn bots in "Instagram jail" to use Kaplan's phrasing.

My research interest centers on the contradictions and double standards in Meta's platform governance that arise from its ambiguous concept of "authenticity," which underpins its opaque and unreliable authenticity governance. The user-led initiative of bot police profiles involved creating new accounts on the platform to establish an Instagram bot police.These profiles detected and documented information about fame-enhancing botting and porn bots. Botting refers to repetitive and quantitative posting, messaging and engaging on social media platforms and involves, in this specific case, fame-enhancing bots to automate posting and engagement. After detection the bot police profiles ensnared the bots to interact with them, publicly pilloried their usernames, and flagged and reported them. In this regard, the bot police enforced the platform's inherent peer surveillance and its call for mutual moderation. The limited possibilities for action led the bot police to elevate themselves above other Instagram users and to engage in digital vigilantism.

In this study, I examine two distinct bot policing methods that are effective in detecting botting and reporting porn bots. This contribution explores how the

platform shapes and influences user agency by investigating why and how user-led bot policing contributes to Instagram's authenticity governance, ensuring alignment with the platform's and bot police's definitions of "authentic" behavior. A comprehensive analysis of user-led bot policing, employing qualitative visual and textual analysis as well as thick description, revealed more intricate dynamics than simply ordinary users reporting botting and porn bots on Instagram.[2] These dynamics highlighted broader power imbalances between the platform and its users and the challenges in regulating user behavior within a privatized, capitalist, public, online environment.

In the first section, I review the challenges posed by "inauthentic" behavior to Instagram's platform moderation and its current implementation. Authenticity governance on Instagram is two-fold: on one hand, the platform often deploys automated governance systems to moderate "inauthentic" behavior. On the other hand, enforced by specific features, the platform allows for peer surveillance and mutual moderation. The following section describes two types of bots that the platform and the bot police consider "inauthentic," which represent the targets of the bot police. It discusses the concept of "authenticity" in relation to botting and the challenges and complexities of detecting fame-enhancing and porn bots. It also includes a literature review on coordinated behavior and the integration of users into platform governance through the flagging feature. The third section introduces the methodology before I analyze two case studies of user-led bot policing approaches. I draw on Tarleton Gillespie's idea of platforms as the "custodians of the Internet" to reveal the bot police's self-understanding as custodians of "authenticity" on Instagram.[3] At the same time, I show that they act as vigilantes, potentially harming other users. My analysis further reveals the mechanisms of peer surveillance and the complex power asymmetries between the platform and its users and between users and other users.

This article adds to a growing body of literature that explores the agency of users in social media platform governance by studying the unexamined practice of

user-led bot policing through this lens. The research contributes to interdisciplinary research on platform culture, bot detection, and platform governance. It is a starting point for research into user-led policing and participation in authenticity governance on Instagram.

Peer Surveillance and Mutual Moderation Inherent to Instagram's Authenticity Governance

Various intentions for engaging in so-called "inauthentic" and unwanted behaviors challenge the moderation of social media platforms. Tarleton Gillespie discusses platform moderation as an integral and essential element of digital platforms, although one that they "take on reluctantly."[4] The outcome of moderating user-generated content, behavior, and engagement of users results in platforms being obliged to moderate, which is why Gillespie conceptualizes them as custodians of the Internet. Platform companies employ different governance strategies, which Thomas Poell and colleagues summarize as regulation, curation, and moderation.[5] One crucial strategy of platform governance is content moderation, which is complex and aligns with the platforms' diverse interests. As part of Meta, Instagram shares its resources for content moderation with other platforms of the parent company, such as Facebook.[6] Instagram promotes its platform governance aimed at undesirable, harmful, or illegal content and behaviors. Meta developed a complex system to manage this content, as Mark Zuckerberg noted in retrospect, "partly in response to societal and political pressure."[7] From 2016 to 2025, a independent third-party fact-checking program was part of Meta's content moderation. Initially, the program was introduced during a time of "heightened concern about information integrity coinciding with the [first] election of Donald Trump as US president," amid heated discussions about how social media platforms foster the spread of misinformation and disinformation. Ironically, Meta

has canceled independent fact-checking due to “too much censorship” and has claimed its “fundamental commitment to free expression.”[8] The press surrounding Zuckerberg’s announcement of the cancellation connects this step to an attempt to improve relations with the recently inaugurated for a second term, US President Trump.[9] With this rollback on moderation, Meta withdraws from responsibilities and places them on their users. The shift to community notes exemplifies how a platform’s governance is influenced by its political and economic interests.

Gillespie observed in his 2018 book on content moderation that “most [platforms] would prefer if either the community could police itself or, even better, users never posted objectionable content in the first place.”[10] Specific mechanisms for mutual moderation have long been part of Instagram’s platform culture, with specific affordances implemented in its interface. The report feature enabled a form of peer surveillance and the development of flagging practices. According to media scholar Oliver Leistert, peer surveillance is already inherent to social media platforms because it invites users to constantly express themselves and comment on others.[11] This aligns with Nicole Cohen’s findings, which outline a political economy of Facebook through the valorization of surveillance and “the unpaid labor of producer-consumers that facilitates this surveillance.”[12] The transition between communicating with one another and monitoring each other is particularly fluid on social media platforms. Leistert defines peer or lateral surveillance as the mutual observation and monitoring of one another, referring to Mark Andrejevic.[13] The fundamental basis of the surveillance paradigm of commercial capitalist platforms is the valorization of the expressions and affects of users, whose likes, shares, and other clicks are analogous to unpaid labor. Given the economic value of user engagement and the effect of producing marketable data for advertisers, it is logical that the platforms are interested in their understanding of “authenticity.” Peer surveillance culminates in the mutual tracking of unwanted content and behavior on the platform, ultimately benefiting its owners.

However, moderating social media platforms involves more than just curating user-generated content. Sarah Myers West has noted that discussions about content moderation have mainly concentrated on the content itself. As platforms have become central private infrastructures for our work, lives, and communities, moderation debates should broaden to encompass the effects of moderating user behavior overall.[14] The moderation of engagement is another crucial aspect of platform governance. Johan Lindquist and Esther Weltevrede define this specific platform moderation as regulating "the boundary between 'authentic' and 'inauthentic' account behavior."[15] Consequently, platforms create systematic and often automated governance systems to moderate "inauthentic" behavior. Wendy Chun refers to this as "algorithmic authenticity," highlighting how the platform's methods "reveal the ways in which users are validated and authenticated by network algorithms."[16] Lindquist and Weltevrede term this platform moderation "authenticity governance" and ground their study in issues of misinformation and disinformation, tackling the metrification, monetization, and political influence of social media. Social media platforms have shifted toward authenticity governance, revising their policies and moderation systems to regulate "inauthentic" behavior. The correlation of terms like "inauthenticity" and "fake engagement" embodies *platform paternalism* and scrutinizes the institutionalized power imbalance between platform companies and their users.[17] Through their regulatory authority, platforms determine what constitutes "acceptable" and "authentic" engagement, influenced by their capitalist and heteronormative logic.[18] Ariadna Matamoroz-Fernández, Louisa Bartolo, and Betsy Alpert argue that platforms have "constructed the notion of 'inauthentic behavior' as a flexible category" to allow for definitions that serve the platform's interests.[19] Users must continually negotiate the platform's definition of "authentic" behavior.[20]

“(In)Authentic” Instagram Bots

At this point, we must distinguish between two types of bots considered “inauthentic” by the bot police, and which are featured in the two case studies. Both bot types challenge the dichotomies between human authenticity and non-human artificiality on social media platforms.[21] The first case study of user-led bot policing targets fame-enhancing bots, which are automated or semi-automated accounts on Instagram. Fame-enhancing bots repetitively and quantitatively post and engage with other users to provoke reciprocal engagement and, as a result, improve engagement metrics, which has a positive effect on the account’s visibility and reach. Therefore, fame-enhancing botting is a strategy users employ as a response to the platform’s governance systems, which demand “an artisanal non-instrumental approach to cultural production”[22] while simultaneously rewarding the opposite behavior with visibility.[23] To be recognized as human users and avoid detection by the platform’s governance algorithms, fame-enhancing bots are programmed to mimic “human” behavior by sharing content and performing interactions such as liking, following, and commenting within a specific and predetermined timeframe and quantity.[24] Automatically performed interactions are complex to distinguish from those executed “manually” by human users, making them complex to detect. I term these automation services “fame-enhancing bots,” their use “botting,” and the respective user a “botter.” The obfuscated nature of botting and the fake profiles designed to circumvent detection pose challenges for both platforms and users in managing their presence. Some Instagram users consider automated engagement and content production as fake engagement, misinformation, and cheating, perceiving them as “inauthentic” behavior. Media coverage surrounding the bot service Instagress’s shutdown in 2017 even characterized the use of botting services as shameless moral deviance and criminality.[25] In examining the Instagress shutdown, Caitlin Petre, Brooke Duffy, and Emily Hund found that the discourse around it invoked “authenticity” concerning audiences: “Specifically,

the notion of 'real follower' was used to distinguish human-run accounts from automated ones (i.e., bots)."[26] Although the metrification and monetization of social media engagement exemplify the need to move beyond the dualisms of fake vs. real,[27] "inauthentic" behavior has often been associated with malicious conduct and content.[28] However, it is essential to differentiate among various forms of "inauthentic" behavior, as it is not always malicious or intended to mislead. Following Matamoroz-Fernández, Bartolo, and Alpert, I argue that the use of botting services as a strategy in the visibility game on Instagram can instead be conceptualized as defiance against platform power, not aimed at misleading or harming fellow users.[29] In many cases, botters are ordinary users seeking to challenge the opaque system of Instagram's algorithmic ranking.

By contrast, such an understanding of botting as a defiance of platform power does not extend to fake profiles or porn bots, which are the main targets of the bot police discussed in the second case study. Elena Pilipets, Sofia Caldeira, and Ana Marta Flores point out that "porn bots operate via sexual solicitation to capture attention."[30] The authors discuss Instagram's "mechanisms of computerized tracking," which enable social automation on the platform, referring to Agre's work on the capture model, which "has deep roots in the practices of applied computing through which human activities are systematically reorganized to allow computers to track them in real time."[31] Like fame-enhancing bots, these bots deploy circumvention tactics to avoid detection by sometimes running fake profiles that mimic existing human users.[32] Contemporary methods of detecting pornography involve human moderation and computer vision algorithms trained on extensive databases of nude images. Meta uses AI systems to assess whether posts on their platforms contain nudity or sexual activity before deciding whether to remove the content.[33] Pilipets and her colleagues found that their policies define sexual solicitation as a "'suggestive' post combined with a request for interaction," concluding that "it is a difficult playground for both human users and porn bots to navigate, for engagement in the form of clicks implies both the promise of attention and the risk of being de-platformed." Porn bots are used to

capture sensitive data, spread malware, or entice users to click on links in their bios. These links often redirect to questionable dating or porn websites behind costly paywalls.[34] Social media platforms like Instagram provide the infrastructure for spreading spam, scams, misinformation, and disinformation. A study by the European non-profit organization AI Forensics on pornographic ads revealed Meta's double standards in moderating pornographic visuals, which are removed when uploaded by a user but promoted when uploaded as a paid advertisement.[35] These double standards negatively affect human content creators more than porn bots.[36] Research into Instagram's platform governance has revealed serious effects on its users.[37]

In addition to overseeing content, Meta introduced its "coordinated inauthentic behavior" (CIB) policies in 2017 and released its findings on CIB in late 2018. Instagram announced the implementation of machine learning tools to identify accounts deploying third-party services for "inauthentic" likes, follows, or comments and to eliminate "inauthentic" activities. Furthermore, they indicate that continued use of third-party services may result in restricted Instagram usage.[38] Coordinated behavior has been examined in relation to disinformation and the manipulation of public opinion in computational social science.[39] Further studies on users' involvement in platform governance focus on the labor and impact of volunteer moderators across various social media platforms.[40]

So far, Meta grants a strongly defined and restricted form of agency and power to act against "inauthentic" behavior to users of their services. The possible sanctions that users can impose against unwanted behavior are limited and built into the interface design. Research on specific features such as flagging has addressed the involvement of users in the platform governance process. Platforms appear to have shifted agency to users through the flagging feature, empowering them to report seemingly inappropriate activity and monitor one another.[41] Kate Crawford and Tarleton Gillespie assert that flags are integrated as a key component of platform governance devolved to users "as a mechanism to

elicit and distribute user labor—users as a volunteer corps of regulators."[42] Users do not receive any education or training, which allows for speculation and interpretation and places some users above others without providing a defined framework. Consequently, there is a significant risk that users will act based on their personal discretion and interests. Crawford and Gillespie have described instances such as gamified flagging among friends.[43] Flagged posts and profiles are frequently reviewed and assigned to underpaid workers with unclear review processes.[44] As Myers West points out, user flagging remains crucial for identifying content that requires removal, thereby affecting the effectiveness of content moderation.[45] The flagging process and subsequent actions taken on a flagged item are opaque and lack transparency, leading to a contentious debate in the literature. Some view it as a vital mechanism in content moderation and an effective strategy to combat misinformation.[46] Others underscore the threats of misuse for specific communities. Carolina Are's study illustrates that flagging is used maliciously on platforms like Instagram and TikTok as a silencing tactic, resulting in de-platforming that harms vulnerable and marginalized communities.[47]

However, research has yet to examine the role of users in governing botting with fame-enhancing bots and porn bots on Instagram. In analyzing the case studies, I will closely investigate the agency of the bot police and the complex power dynamics at play. The preceding discussion indicates that user-led bot policing on Instagram appears to fulfill Meta's aspirations: Instagram users who voluntarily detect, flag, and report botting and porn bots. The following section examines why and how user-led bot policing contributes to Instagram's authenticity governance to ensure compliance with the platform's and the bot police's understanding of "authentic" behavior.

Authenticity Governance Through User-Led Bot Policing

To ground the preceding arguments empirically and reveal the power dynamics, this section turns to two bot policing profiles as case studies that employ different strategies and vary in their approaches and actions against botting and porn bots. Examining these two profiles aims to showcase diverse examples of bot policing targeting two distinct types of bots. The profiles were selected for several reasons: the first case study of bot policing focused exclusively on combating automated interactions from botters deploying fame-enhancing bots, offering valuable insights into botting practices. The second case study featured both a German and a US bot police profile. The German bot police profile was chosen from a coordinated group of bot police officers, potentially initiated in response to a surge of spam bots in early 2019.[48] Most German bot policing profiles were primarily active in 2019, whereas the US bot police serve as an active example with a notable number of followers. The German bot police group has appeared inactive since 2020, having published their last posts that year. The profiles did not provide reasoning for their cessation, but before this, the movement announced its aim to eliminate bots by setting a deadline of 2020, using the hashtag #nobotsby2020. Reaching their time limit could explain their abandonment of this goal. The US bot police profile also employed the hashtag #nobotsby2020. Unlike the German bot police, this profile remains active as of August 2024. One of their story highlight folders proudly features popular Instagrammers with over two million followers following the bot police profile. Their support and their substantial follower count are sources of legitimacy for them.

These case studies offer a distinct, publicly observable perspective on various user-led practices aimed at countering botting and porn bots. A qualitative visual and textual analysis of the profiles was carried out to understand user-led bot

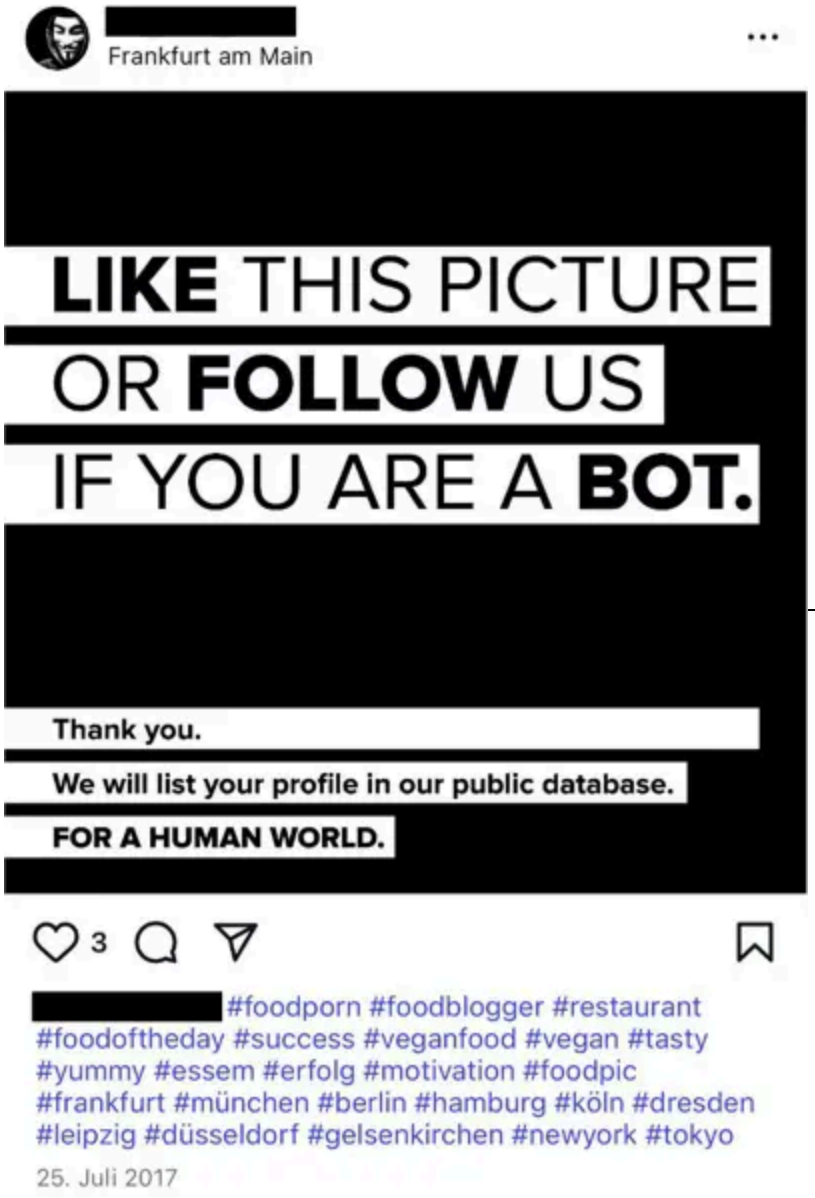
policing on Instagram, focusing on its objectives and strategies. This method considers the co-constitutive relationship between Instagram accounts and the components of the profile's posts, including content, hashtags, captions, comments, and interaction metrics.[49] Instagram's built-in search functions enable users to search for usernames and hashtags; hashtags serve both as an "initial point of departure for studying activity on Instagram" and as tools to locate and trace data.[50] Data for this study was manually collected through the Instagram user interface and preserved using screen capture. This approach ensured that the relevance of the data was assessed while minimizing data collection. The bot policing profiles were accessed in October 2023, March 2024, and August 2024. Data was gathered without consent but was re-verified to confirm it remained public before submission. The bot policing profiles do not disclose the identity of account holders. To protect Instagram users' privacy, all sensitive data or identifiable information of targeted botters has been excluded or anonymized. The study's limitations include a small qualitative sample of case studies and restricted access to data. The obfuscated nature of the research subject presents significant challenges for research in this field, as botting and porn bots operate at different and sometimes impossible levels of observability.

Furthermore, I used Clifford Geertz's concept of "thick description" to examine the culture of the bot police.[51] The thick description specifies details about their social structures, actions, and meanings, providing in-depth knowledge about how the bot police "feel, think, imagine, and perceive their world."[52] In this regard, thick descriptions help in understanding the context of the bot police's actions, "stat[ing] the intentions and meanings that organize the actions," "trac[ing] the evolution and development of the act," and "present[ing] the actions as a text that can be interpreted."[52] In this way, the methodological approach uncovers the power dynamics between the platform, its users, and their peers. The analysis emphasizes the complexity of moderation and authenticity governance through the interconnectedness of "authenticity" and artificiality on social media platforms. With case studies based in Germany and the USA, and as a researcher

living and working in Germany, we learn about user-led bot policing from a Western and democratic perspective within a capitalist society. User-led bot policing may differ in non-Western and non-democratic environments, operating in different languages and online cultures with distinct platform values, potentially targeting various forms of “inauthentic” behavior on Instagram.

We Find Fame-Enhancing Bots—Case 1

The first case study examines actions taken against fame-enhancing bots. Botting, used as a visibility-enhancing tactic, is an example of users opposing the platform’s authenticity governance as defiance of platform power while simultaneously redefining the boundaries of “authenticity.” Visibility-enhancing tactics through automation are condemned as “contamination,” “crime/moral deviance,” and “cheating.”[54] These metaphors reveal the double moral standards upheld by social media platforms that deploy automation tools for content moderation and authenticity governance. The first bot policing profile aimed to expose the automated interactions of botters. As fame-enhancing bots are complex to detect, the bot policing profile devised a new strategy by posting content that non-botters were unlikely to engage with, in order to phish for and actively ensnare automated engagement (fig. 1).



https://csalateral.org/wp/wp-content/uploads/2026/01/Call-the-Bot-Police_fig-1-Nathalie-Schafer.jpg>

Figure 1. Screenshot of a post by the bot policing profile.

At the same time, they ensured that non-botters would refrain from interacting with them by threatening public denunciation. In these phishing-for-engagement posts, they tagged various cities in Germany and included popular hashtags. The hashtags and location tags served as lures for botters since fame-enhancing bots must target specific hashtags or locations. Analyzing the interactions with the

phishing posts made it impossible to verify whether a human or a bot had liked the content. Firstly, the profile assumed that every like was automatically generated and showed no indication of verifying in any manner whether they were automated likes. Secondly, as mentioned earlier, this raises questions about the status of likes and “authenticity” within the context of “algorithmic authenticity” in metrified and monetized social media environments. The dichotomies between “inauthentic” and “authentic,” fake and real, and human and artificial are neither sustainable nor useful because engagement on social media platforms inherently crosses these boundaries.

The comments on the posts were mainly vague and broad, often consisting of just a single emoji. One comment appeared to be made by a fame-enhancing bot, stating: “Thumbs up, beautiful photos☺️ I hope you like my profile, too. Best wishes, [botter’s name and affiliation].” Comments can only be tentatively attributed to fame-enhancing bots due to the lack of context, as seen in this instance. Instagram’s user interface does not provide features or tools for users to filter automated from manual engagement. Because of this limitation, this bot policing profile developed alternative methods of peer surveillance and mutual moderation. At the same time, they positioned themselves above other users, risking trapping potentially “real” and “authentic” engagement. By threatening other users with public denunciation, they shifted the power balance hierarchically downward.

<

https://csalateral.org/wp/wp-content/uploads/2026/01/Call-the-Bot-Police_fig-2-Nathalie-Schafer.jpg>

Figure 2. Screenshot of bot policing profile with profile picture and screenshots of the construction companies' interaction with them.

While likes and comments constitute key modes of engagement, the account's profile picture further contributes to their self-understanding as activists. The profile picture displayed a figure in a hoodie and a Guy Fawkes mask associated with the hacktivist movement Anonymous(fig. 2).[55] According to Christian Fuchs, Anonymous is a

social movement and anti-movement; it is collective political action based on a shared identification with some basic values (such [as] civil liberties and freedom of the Internet) that results in protest practices online and offline against adversaries, and at the same time for many of those engaging on Anonymous platforms individual play and entertainment.[56]

What platform values does this profile likely align with, particularly on Instagram? Social media platforms are sites where values are "expressed, contested, and diffused." Platform values refer to the "underlying principles governing and expressed through social media."[57] A platform's cultural values can be discerned by examining community norms, as illustrated through established "authenticity" practices or governing documents. Policy documents detail how various actors should behave and reveal the core principles of the platform company's values. The user's values are reflected through practices, usage patterns, and user-generated content. Key values on Instagram include engagement and "authenticity." Hallinan and colleagues further emphasize that social media platforms "have simultaneously been credited with . . . championing authentic self-expression and incentivizing fake interactions."[58] As previously noted, "authenticity" on social media is a vague term with varying interpretations depending on perspective.[59] Instagram articulates "authenticity" both as an informational mechanism and as a normative framework for user behavior. On one side, the platform adopts an informative understanding of the term, addressing the necessity of authentication and the importance of verification badges for specific users, such as businesses and public figures. On the other side, Instagram employs the term to indicate and communicate its expectations

regarding how users should engage and conduct themselves on the platform, framing these expectations as essential for safety, trust, and security. Meanwhile, economic interests in user data and engagement drive the platform's expectation of "authentic" behavior. Instagram's definition of "authenticity" remains deliberately vague, allowing it to be flexibly adapted to changing objectives. The fact that Instagram's algorithmic ranking systems reward botting behavior suggests that they also benefit from the traffic generated by such engagement.

By contrast, users commonly interpret "authenticity" as being "oneself," emphasizing forms of "realistic" self-expression. This notion of individual "authenticity" remains ambiguous in itself. Instagram posts related to "authenticity" tend to address topics such as being yourself, self-care, and self-improvement. Ironically, the intended thought-provoking content is often connected to images containing highly professional composition, editing, and carefully curated backgrounds. It raises questions about the construction of identity online which is more complex than the expectation of "being oneself" online.[60] This notion of "authentic" likes is also referred to as "organic" or "genuine," implying that a liking user has a true interest in a post due to its content or their interest in the profile that posted it.[61] This simplistic and one-sided perspective fails to recognize the platform's culture and its inherent visibility game within the attention economy, overlooking the various motivations for engaging with content posts or comments, such as influencer pods.

The bot policing profile's reference to Anonymous highlights their identification with the hacker group that "constitute[s] a global, Internet-based social movement of primarily young and dissatisfied people who have decided to follow their own strategy to oppose established global governance and order."[62] The profile materializes this definition by actively running the account as a site of bot policing. In this case, the bot policing profile focused on authenticity governance and upholding their interpretation of platform values by bringing two approaches against botters into practice.

Firstly, they followed all profiles that interacted with them in any way, hereafter referred to as "followees." One can assume that all their followees engaged in botting at some point. Followee numbers fluctuate over time, ranging between 280 (current count, January 2026) and 310 (at the time of first data collection in February 2023). Given the brief active period in July and August 2017 and the resulting follower reach, the number of botters interacting with them is noteworthy. This is significant both because it allows us to draw rare conclusions about the notable automated engagement during that time, and because it was the only opportunity thus far to observe engagement conducted by fame-enhancing bots since I began researching them in 2018.

Secondly, they shared screenshots of their notification of interactions list, which included a notification of a construction company that liked their posts (fig. 2). The bot policing profile commented: "Construction company uses social bots 😠." The construction company again liked the screenshot, which prompted the bot police to post another screenshot, stating that illegal social bots would be challenging to regulate. This example pushed the mechanisms of peer surveillance to the extreme. Beyond their trap for automated engagement, they captured and published information about engagement metrics intended solely for their use. The notification feature exemplifies the metrification of engagement and the valorization of peer surveillance par excellence.

The bot policing profile, alongside other users engaging with the denunciation posts, viewed botting as a moral deviance and accused botters of lacking honor. Consequently, the bot policing profile devised its own strategy for authenticity governance, targeting botters by publicly shaming everyone interacting with their account and content. Favarel-Garrigues and colleagues describe this phenomenon as digital vigilantism, characterized by "a basic principle of 'naming and shaming', or through a 'weaponisation of visibility', that is sharing the target's personal details by publishing/distributing them on public sites ('doxing')."[63] In this scenario, the personal details included the botter's profile name and picture.

As the reasoning goes, the loss of honor typically leads to feelings of shame. Disgrace has likely existed as long as humanity itself and was a common formal punishment across virtually every culture. It was also one of the most prevalent punishments during the Catholic Church's Inquisition. These honor-related punishments were often executed before large audiences at the pillory. "Those who had forfeited their honor were subjected to public, often merciless, ridicule,"[64] which mirrors the strategy employed by this bot policing profile. Yet, there is a certain irony in that visibility was a central element of such vigilante campaigns, which have evolved into digital vigilantism. Applying this form of punishment to botters is highly problematic for at least three reasons: First, the bot policing profile lacked the authority to prosecute, judge, or punish botting. Second, botters seek visibility in defiance of Instagram's authenticity governance without the intent to harm, mislead, or endanger the community, which does not provide a solid basis for prosecution and mutual moderation. Third, pillorying is generally not an appropriate punitive measure for "inauthentic" behavior on Instagram, even though it has resurfaced and become a recurring phenomenon in online environments.

In this case study, the user-led bot policing action was restricted to detecting botting and publicly denouncing and pillorying botters. None of the posts or comments indicated that this bot policing profile flagged the trapped accounts. Although this public profile has not been active since 2017, it served as a unique archive and record of bot actions to enhance fame, considering the nearly three hundred accounts it follows.

Call the Bot Police—Case 2

The second case study of user-led policing focused on porn bots and featured profiles known as bot police, operating across various countries and in different languages. Two bot police accounts, one based in Germany and the other in the

USA, were selected to illustrate and analyze a different form of user-led bot policing within authenticity governance. Both profile names included the terms “bot” and “police,” accompanied by profile pictures depicting police officers. They distinguished between two types of porn bots. According to their observations, the first type typically had only one post that didn’t tag other users, yet they would regularly delete older posts and add new ones. The commentary function was mainly disabled. The German bot police profile reported that these porn bots present themselves as female users, and in some cases, offer nude content for sale.

The second type of bots included profiles that promoted links in their bios leading to spam or porn websites and those that commented in a “stupid” and “annoying” way which included using explicit pornographic language and suggestive emojis related to sexual themes. They contained “stupid” story highlights and bios with sex-related information. Another distinguishing feature of the bots identified by the German bot police profile was their use of formal German language forms to address other users, as opposed to the everyday language typically used on the platform. This formality likely stemmed from their reliance on translation tools like Google Translate, which may not account for context or the correct forms of address. While certain features overlapped and the bot police’s typology was not detailed enough to definitively distinguish two divers types, both refer to pornographic spam attempting to spread malware, monetize nude content, or entice users to click on links in their bios. Both bot profiles generated and shared knowledge about their targets and encouraged their followers to observe their peers, identify porn bots, and collaboratively moderate the actions of other Instagram users.

The profiles’ objectives focused on policing porn bots. Addressing users’ moderation efforts, Gillespie refers to them as amateur editors and police: “By shifting some of the labor of moderation, through flagging, platforms deputize users as amateur editors and police.”[65] It is no coincidence that Instagram users

engaging in user-led bot policing refer to themselves as “bot police.” Traditionally, policing was defined as “a system to enable the well-being of individuals and the growth and welfare of the community.” [66] The term “police” has its roots in Classical Greek (politeia) and pertains to “all matters affecting the survival and welfare of the city.”[67] The shared linguistic root might suggest that only the institution of the police would be in charge of “policing.” Clive Emsley’s *A Short History of Police and Policing* illustrates a more complex development of the term “police,” highlighting how policing practices have been assigned to civil members of communities throughout history in order to enforce social norms. The bot police assumed the role of Instagram police, which does not align with the concept of police as an institution, recognized as a state organ and part of its executive authority. In the more restricted understanding, “police refers to the state authorities whose task it is to avert dangers to the individual and the general public, to protect public order and security and to prosecute criminal acts.”[68] The German bot police profile was labeled a “police station,” and the account holder self-identified as a voluntary bot police officer. This account was part of a group of German profiles with usernames referencing bot police or bot police officers. Several users with bot police profiles organized and coordinated themselves, utilizing a Discord server for a “united bot police” on Instagram. To gain access to this server and the group, interested future “police officers” could contact the administrators. This coordinated bot policing represented an attempt to institutionalize peer surveillance and mutual moderation of “inauthentic” behavior.

The case of the US bot police profile was somewhat unique. This profile operated as a single police officer without gathering additional police officers, though they managed their own Discord server. The US bot police profile also ran another account called FBBI, which stood for “Federal Bureau of Bot Investigation,” parodying the FBI of the United States. It served an entertaining purpose by sharing bot memes with the bot police profile’s large community of over 280,000 followers. They leveraged their reach to recruit their followers as voluntary

“auxiliary policemen” to aid them in policing porn bots through coordinated collective flagging. Daniel Trottier described this phenomenon as digital vigilantism, where citizens were collectively offended by other users’ actions and retaliated through coordinated responses on digital media platforms, including mobile devices and social media.[69] The German bot police profile created a highlights folder with instructions on removing porn bots. The bot police explained how their followers could flag porn bots by showcasing a series of screenshots with instructions. These instructions included where to find the reporting function on Instagram’s interface and how to flag porn bots as spam. Both bot police profiles urged their followers to tag them or send links to porn bots, functioning as collection points for these reports. Subsequently, they published the information and called on their community to use their combined strength to flag and report collectively. This bot policing profile integrated the actions of peer surveillance, shaming, and digital vigilantism. Notably, the bot police highlighted the notification a user receives after flagging a profile and hinted at the platform’s review process. The notification contained information indicating that the platform blocked the account as a result of the bot police’s report. For the German bot police, this notification was a reward for their “successful” intervention against “inauthentic”' behavior. A reel by the US bot policing profile demonstrated the rewarding effect of the action of reporting a porn bot (fig. 3).

bot_police
The Kid LAROI, Justin Bieber · STAY

View profile

Watch on Instagram

View more on Instagram

16,720 likes
bot_police

How I feel after reporting a bot
-

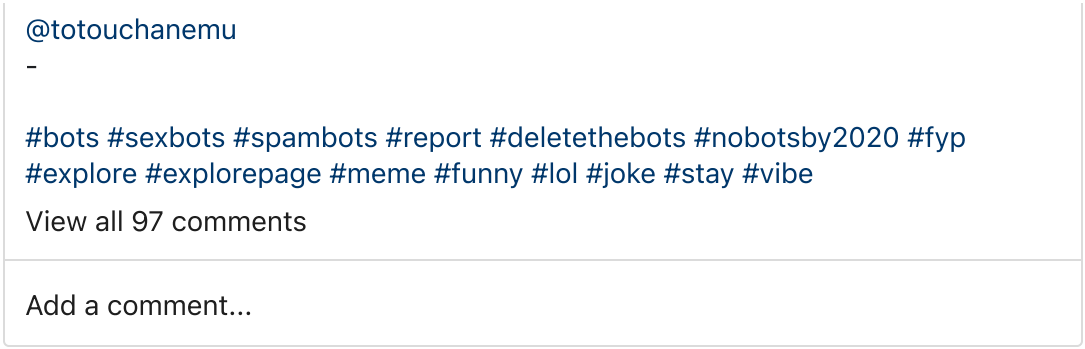

@totouchanemu
-
#bots #sexbots #spambots #report #deletethebots #nobotsby2020 #fyp
#explore #explorepage #meme #funny #lol #joke #stay #vibe
View all 97 comments
Add a comment...

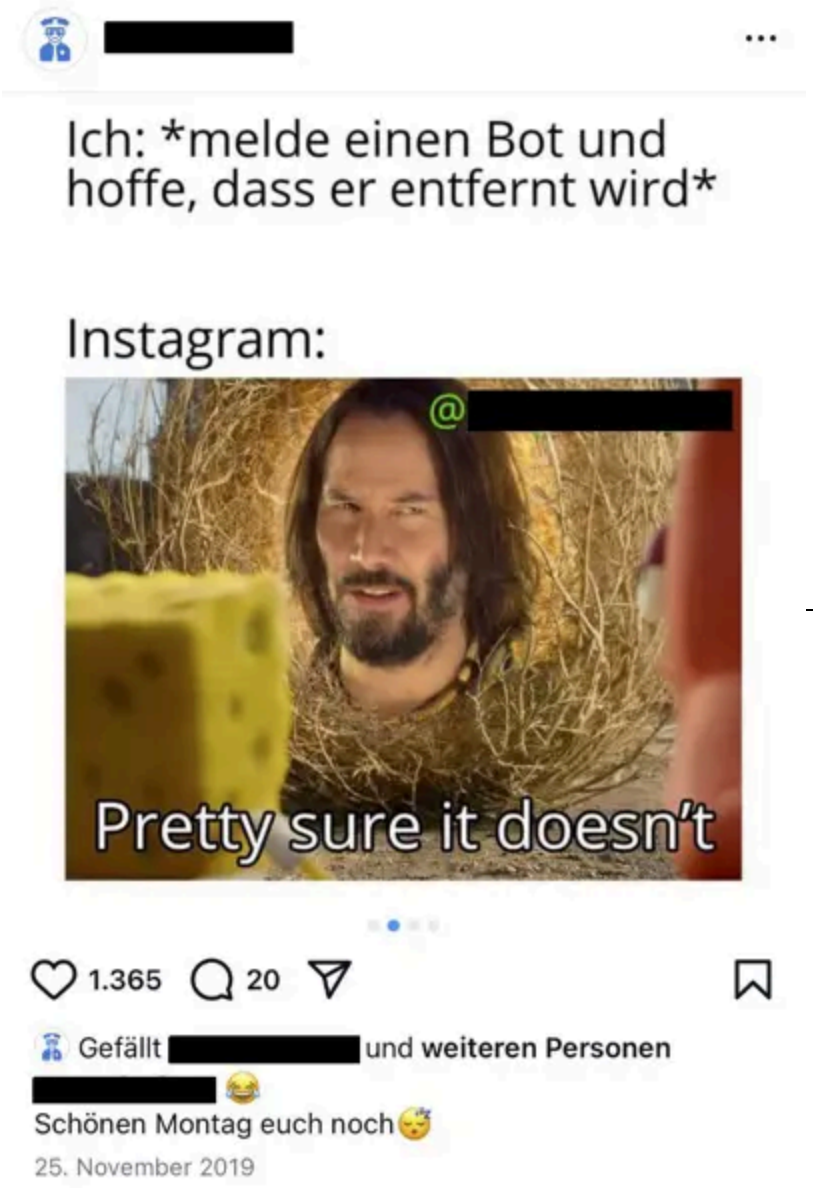
Figure 3. Reel demonstrating the rewarding effect of reporting porn bots.

Apart from that, both bot police profiles also had an entertaining factor for their followers. The German bot police profile entertained its followers with handcrafted memes. Most of their posts were memes that served different purposes. On the one hand, they represented the account holders' identification as a bot police officer who expressed their emotions about their work duties and successes and failures (fig. 4). On the other hand, they functioned as knowledge transfer about the "bot problem" and entertainment for the profile's followers. The memes had similar messages, addressing the vast number of bots on Instagram, the platform company that did not seem to care (fig. 5), and the enormous amount of work for the bot police (fig. 6). The memes showed how the bot police perceived Instagram's limited role in fighting porn bots on the platform, thereby positioning the bot police's own intervention as both necessary and justified.

<

https://csalateral.org/wp/wp-content/uploads/2026/01/Call-the-Bot-Police_fig-4-Nathalie-Schafer.jpg>

Figure 4. German bot policing profile conveying emotions about their work duties: "Me (@[bot police's name]), when I get tagged on a bot again."



<

https://csalateral.org/wp/wp-content/uploads/2026/01/Call-the-Bot-Police_fig-5-Nathalie-Schafer.jpg>

Figure 5. Meme indicating Instagram's carelessness: "Me: *reporting a bot and hoping that it will be banned*."

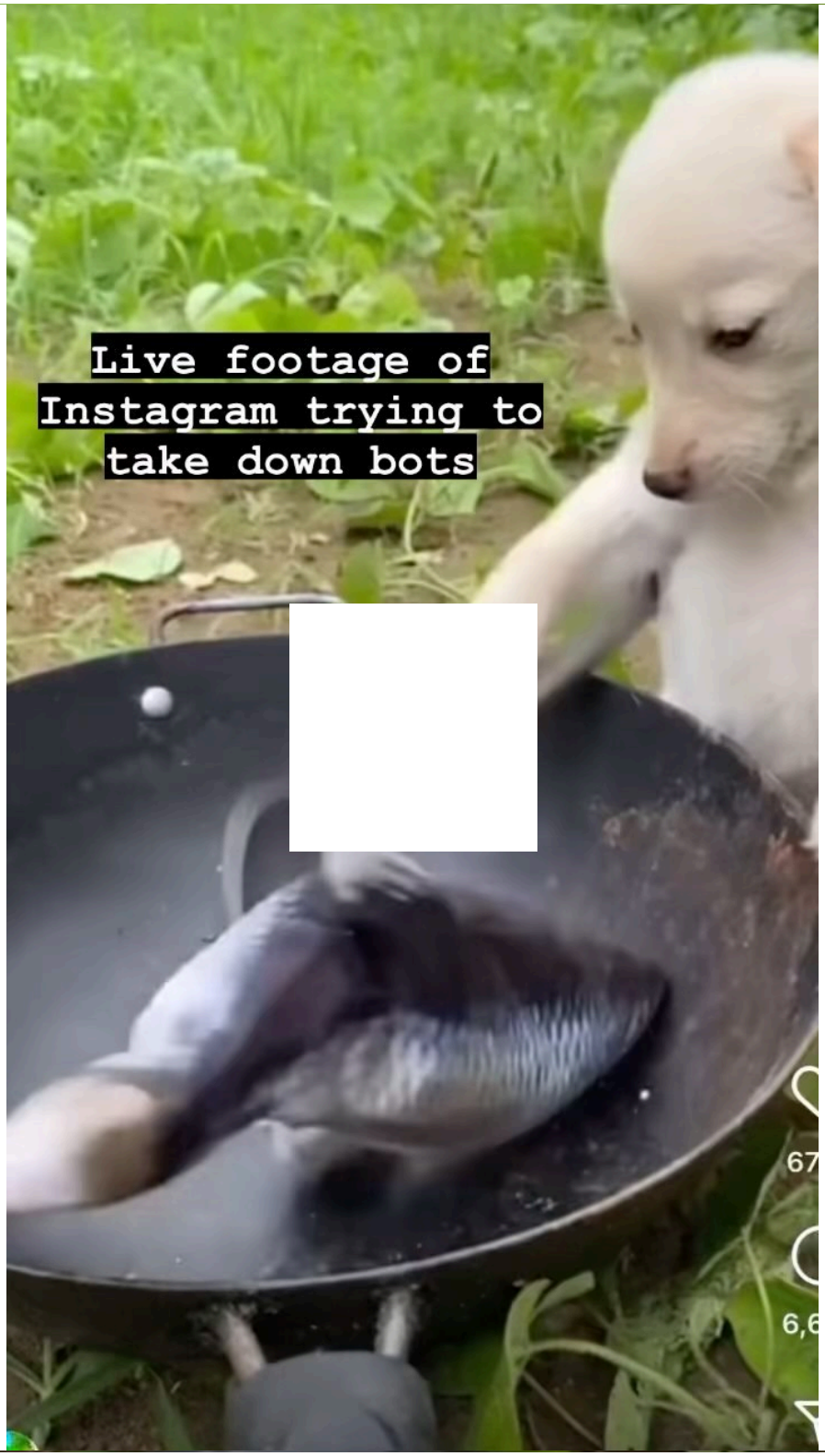
https://csalateral.org/wp/wp-content/uploads/2026/01/Call-the-Bot-Police_fig-6-Nathalie-Schafer.jpg>

Figure 6. Meme showcasing the German bot police's perceived enormous amount of work.

The US bot police posted similar memes and reels on their FBBI and the bot police profile (fig. 9) but, at the same time, communicated their respect for the platform. While the bot police's memes criticize Instagram for its insufficient actions against bots, yet do so in a notably restricted manner. (fig. 7).

bot_police
Original audio

View profile



View more on Instagram

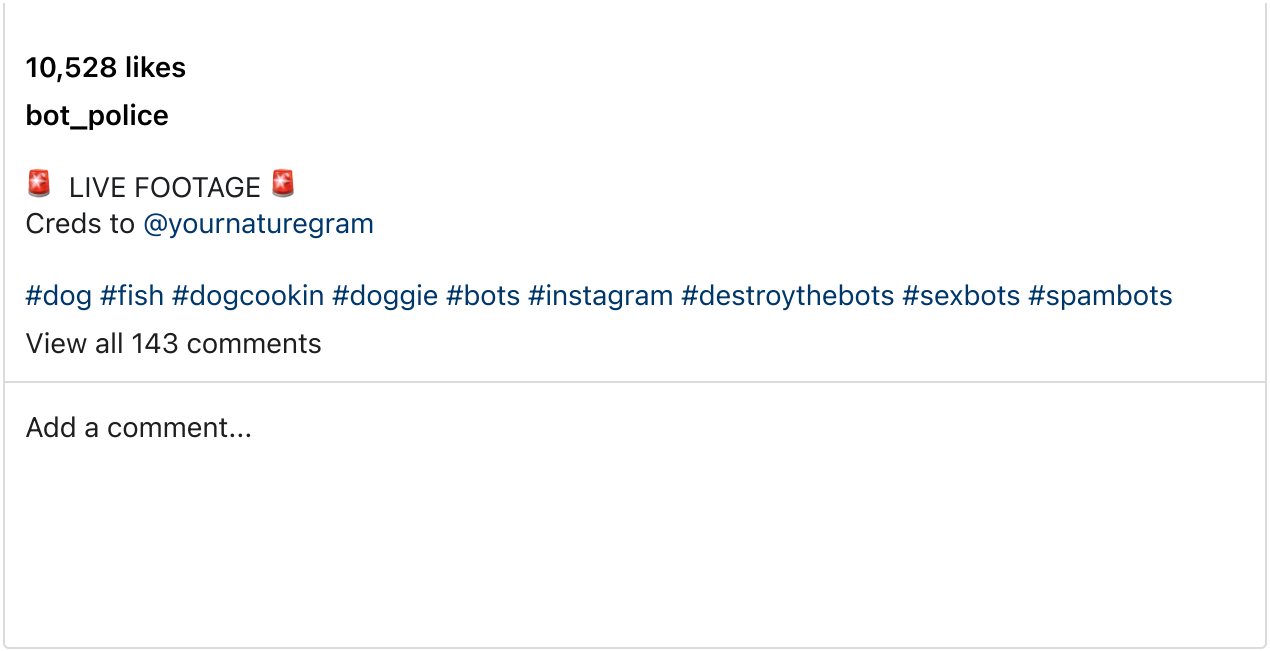


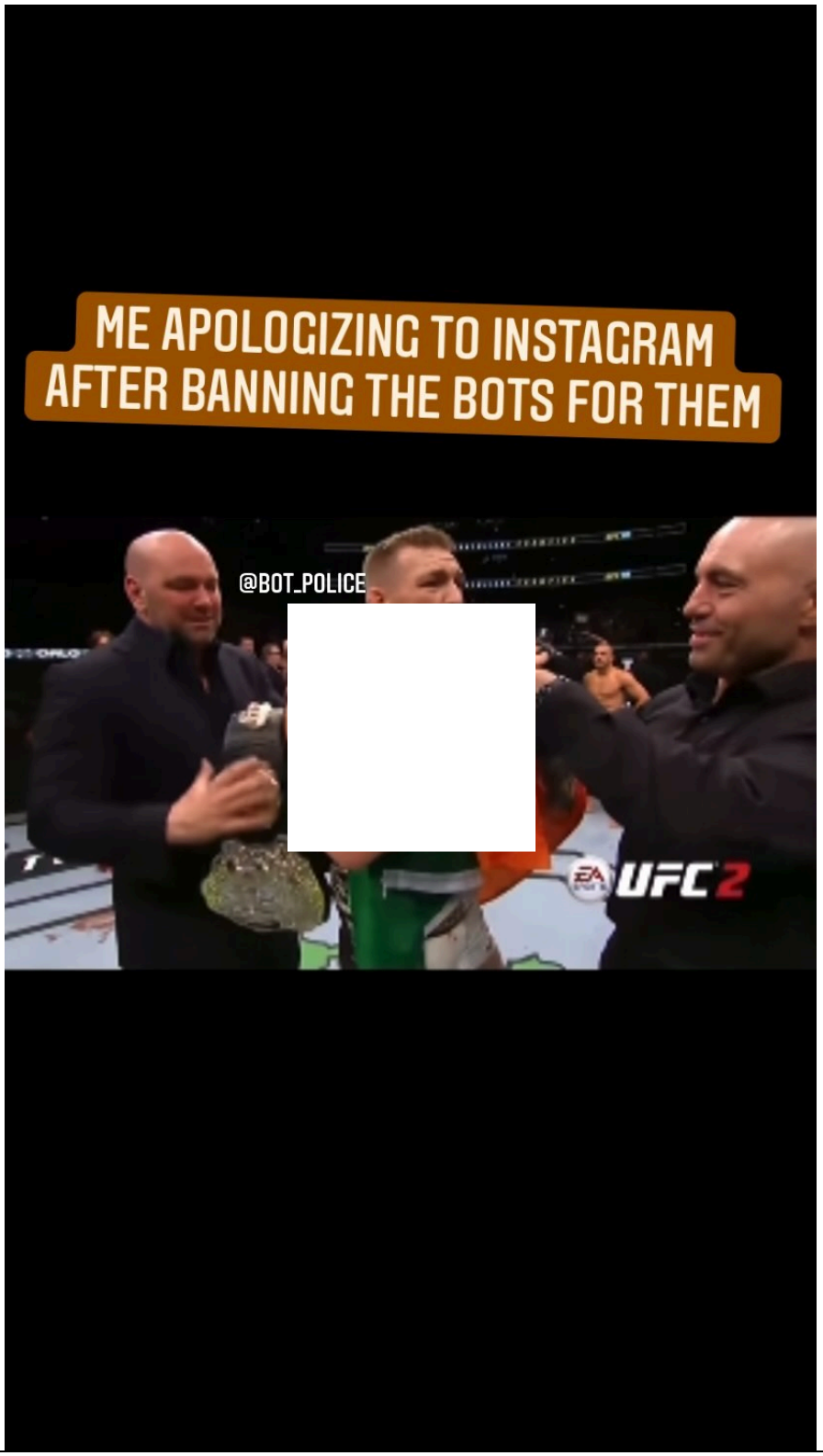
Figure 7. Reel humorously criticizing Instagram's insufficient actions against porn bots.

The criticism was frequently accompanied by appeasing language in the caption, in which the bot police explicitly asked Instagram not to ban it, reflecting an awareness of the risk of platform sanctions and underscoring the asymmetric power relationship between users and the platform (fig. 8). There was no proof that Instagram had banned them. Still, the repeated plea emphasized the dependency and top-down power dynamic between the platform owners and their users. Furthermore, the memes, other content, and comments of the bot police indicated that the platform's actions against reported profiles were opaque, incomprehensible, and unreliable. This argument and the research on the effects of moderation and de-platformization of nudity and sex-positive content following flaggings[70] supports the result of the AI Forensics study that Meta's double standards affect human users rather than porn bots.

https://csalateral.org/wp/wp-content/uploads/2026/01/Call-the-Bot-Police_fig-8-Nathalie-Schafer.jpg>

Figure 8. Bot meme by US bot police profile ironically visualizing Instagram's attitude as "No need to thank me."

bot_police
Original audio
View profile



View more on Instagram

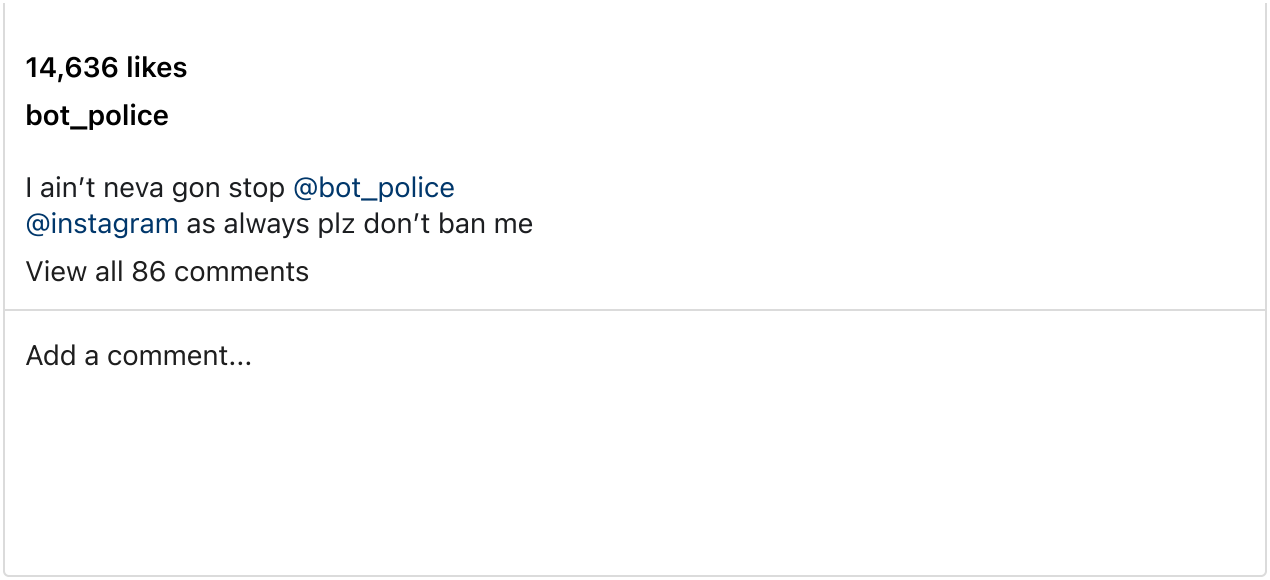
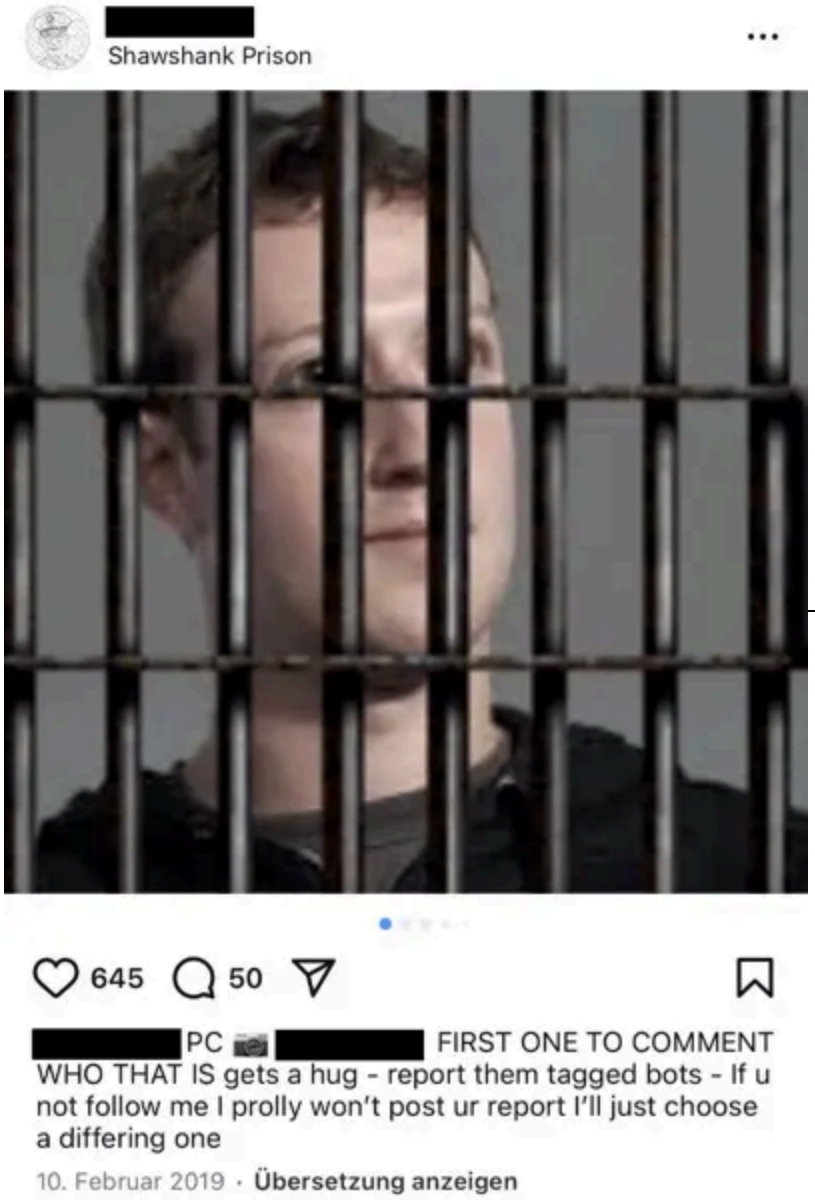


Figure 9. Reel showcasing the US bot police profile's attitude towards Instagram: "I'd like to take this chance to apologize to absolutely nobody."

The US bot police profile used another strategy to entertain their community, provoke interaction, and attract new followers. They regularly posted images of well-known figures with a filter that put them literally behind bars, taking them metaphorically into custody, like in the example with Mark Zuckerberg (fig. 10). In the caption, they challenge their followers to guess and comment on their names. The first correct guess in the comments section was rewarded with visibility by tagging the winner in a following post. The visibility rewards highlight the visibility game that the platform's algorithmic ranking, attention economy, and culture affords. These posts were accompanied by screenshots of comment sections with porn speech or profiles showing nude content of targeted porn bot profiles. The profile's use of entertainment features and memes targeting Instagram exhibited characteristics of pillorying, as their large followership underscores the critical role of publicity and the audience, without which naming and shaming would be ineffective. As del Ama put it: "Public disgrace still has a high news value."[71]


[redacted]
Shawshank Prison
645
50
[redacted] PC [redacted] FIRST ONE TO COMMENT WHO THAT IS gets a hug - report them tagged bots - If u not follow me I prolly won't post ur report I'll just choose a differing one
10. Februar 2019 · Übersetzung anzeigen

https://csalateral.org/wp/wp-content/uploads/2026/01/Call-the-Bot-Police_fig-10-Nathalie-Schafer.jpg>

Figure 10. Exemplary post by US bot police profile of public figures behind bars. In this case, Mark Zuckerberg.

The bot police profiles did receive criticism from other Instagrammers. A user created a profile demanding to stop the user-led bot policing. In their opinion, the bot police were useless, and some bot police officers would use the movement as

a fame-enhancing tactic for their own purposes. Some bot police officers would also take their role too seriously, identifying with a "real" police officer. In the bot police officers' opinion, it would be Instagram's responsibility to moderate bots effectively, but the platform failed since there were still many bots around in an ever-growing number. This critic also addressed some problematic cases where Instagram would ban innocent accounts after flagging. The criticism included that this was an unacceptable injustice and an essential reason for the bot police to find different ways of taking action. The latter suggestion is interesting because it conveys that the bot police critic did not question the bot police's existence in the first place but only their way of operating, which the critic requested they change.

Bot policing is a limited form of authenticity governance at a user level, and the platform's sanctioning actions following flaggings not only serve for "good" but also cause problematic instances. Meta's co-moderation affects marginalized user groups, as in the case of the censorship of two posts on Facebook by an account that belongs to an US couple who identify as transgender and non-binary. They posted topless photos of themselves with their nipples covered. After users flagged them, the platform's automated content moderation system reviewed and removed the posts. Meta restored the posts after the couple opposed the censorship decision.[72] The users who flagged the couple's posts abused the report function to police others' bodies and self-expression. The platform lacks an effective governance measure to define, standardize, and, on this basis, decide what is inappropriate. This lack facilitates and can lead users to harass, discriminate, and censor particular other user groups and identities.[73] As a reaction to the censorship of the two posts, Meta's oversight board recommended revising and modifying their community standards concerning adult nudity and sexual activity to "define clear, objective, rights-respecting criteria . . . so that [nudity] is governed by clear criteria that respect international human rights standards."[74] As I've mentioned above, and recent research shows, Meta still lacks a responsible and respectful handling of censored or shadowbanned users.[75]

Conclusion: The Bot Police as Custodians of “Authenticity” on Instagram

This study investigated the unresearched practice of user-led bot policing against botting and porn bots on Instagram. Two case studies showed various strategies for user participation in Instagram’s authenticity governance. Users’ scope of action was limited to flagging and reporting profiles or content. There was no feature for governing “inauthentic” liking or following behavior. The key findings showed that authenticity governance on Instagram is crucial for some users, motivating those who saw botting as moral deviance to become active. The user-led bot policing profiles developed pillorying and coordinated flagging strategies to participate in Instagram’s authenticity governance. They considered themselves necessary to protect “authenticity” and the platform values and perceived themselves as its custodians. Their perspective of “authentic” behavior as “genuine” or “real” behavior does not account for the hybrid and the entangled nature of human and non-human agents in social media environments. The bot police acted autonomously without having authority from the platform or an institutionalized code of conduct to execute their approaches to authenticity governance. Their digital vigilantism potentially caused harm to other Instagrammers. The pursuit of visibility was not only the reason for botting and deploying porn bots but also a strategy used by bot police to establish their own legitimacy and a motivation for their bot-policing activities. However, any individual user engaging in either botting or bot policing had limited agency, and the case studies demonstrated the imbalanced power relations between users and the platform.

In this respect, the findings revealed complex power asymmetries regarding user-led authenticity governance. Instagram is controlled by its parent company, Meta, which pursues its economic interests and is under political-economic pressure

from politicians and customers. To avoid responsibility, authenticity governance has been shifted to users. The flagging feature is critical to consider, as it grants limited power and agency to users and enforces mechanisms of peer surveillance and mutual moderation, placing certain users above other users. Without clear instructions, a standardized framework, and training, user-led bot policing is inappropriate to ensure "authentic" behavior on the platform. However, "authenticity" itself is a vague concept and a double-standard marketing term, as "inauthentic" behavior is promoted by algorithmic ranking while simultaneously prohibited by Instagram's community guidelines. A bot police force empowered by the platform would have to be trained and act by a code of conduct. This form of authenticity governance would, however, be paid labor, which Instagram already abolished in the case of the independent third-party fact-checkers. How the implementation of community notes on Instagram will affect user-led bot policing remains to be seen. As long as Instagram is a privatized capitalist platform, we will have contradictory authenticity governance, prioritizing political-economic interests over user interests and service for a good and safe societal public space. Future research could address the evolving status and governance of "authenticity" in an AI-driven online environment and the implications of advanced AI-run fame-enhancing and porn bots and research the impact of community notes on user-led bot policing practices.

Notes

1. Joel Kaplan, "More Speech and Fewer Mistakes," Meta, January 7, 2025, https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ < https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/> .

2. Linnea Laestadius, "Instagram," in *The SAGE Handbook of Social Media Research Methods*, ed. Luke Sloan and Anabel Quan-Haase (SAGE Publications Ltd, 2016), 571–92, https://doi.org/10.4135/9781473983847 < https://doi.org/10.4135/9781473983847> ; Clifford Geertz, "Thick Description: Toward an Interpretive Theory of Culture," in *The Interpretation of Culture. Selected Essays* (Basic Books, 1973).

3. Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press, 2019), https://doi.org/10.12987/9780300235029 < https://doi.org/10.12987/9780300235029> . ↩

4. Gillespie, *Custodians of the Internet*, 5. ↩

5. Thomas Poell, David Nieborg, and Brooke Erin Duffy, *Platforms and Cultural Production* (Polity Press, 2022): 7. ↩

6. Tarleton Gillespie et al., "Expanding the Debate About Content Moderation: Scholarly Research Agendas for the Coming Policy Debates," *Internet Policy Review* 9, no. 4 (October 21, 2020): https://doi.org/10.14763/2020.4.1512 < https://doi.org/10.14763/2020.4.1512> . ↩

7. Joel Kaplan, "More Speech and Fewer Mistakes." ↩

8. Kaplan, "More Speech and Fewer Mistakes." ↩

9. Liv McMahon, Zoe Kleinman, and Courtney Subramanian, "Meta To Replace 'Biased' Fact-checkers with Moderation by Users," January 7, 2025, https://www.bbc.com/news/articles/cly74mpy8klo < https://www.bbc.com/news/articles/cly74mpy8klo> ; Casey Newton, "Inside Meta's Dehumanizing New Speech Policies for Trans People," *Platformer*, January 10, 2025, https://www.platformer.news/meta-new-trans-guidelines-hate-speech/?ref=platformer-newsletter < https://www.platformer.news/meta-new-trans-guidelines-hate-speech/?ref=platformer-newsletter> ; Jakob Von Lindern, "Mark Zuckerberg: Er Will Auch Mitspielen," *ZEIT ONLINE*, January 7, 2025, https://www.zeit.de/digital/2025-01/mark-zuckerberg-facebook-instagram-meta-donald-trump < https://www.zeit.de/digital/2025-01/mark-zuckerberg-facebook-instagram-meta-donald-trump> . ↩

10. Gillespie, *Custodians of the Internet*, 6. ↩

11. Oliver Leistert, "Soziale Medien als Technologien der Überwachung und Kontrolle," in *Handbuch Soziale Medien*, ed. Jan- Schmidt and M Taddicken, (Springer Fachmedien Wiesbaden, 2016), 571–92, https://doi.org/10.1007/978-3-658-03895-3_13-1 < https://doi.org/10.1007/978-3-658-03895-3_13-1> . ↩

12. Nicole S. Cohen, "The Valorization Of Surveillance: Towards A Political Economy Of Facebook," *Democratic Communiqué* 22, no. 1 (March 1, 2008): 5. ↩

13. Mark Andrejevic, "The Work of Watching One Another: Lateral Surveillance, Risk, and Governance," *Surveillance & Society* 2, no. 4 (2004): https://doi.org/10.24908/ss.v2i4.3359 < https://doi.org/10.24908/ss.v2i4.3359> . ↩

14. Gillespie et al., "Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates." ↩

15. Johan Lindquist and Esther Weltevrede, "Authenticity Governance and the Market for Social Media Engagements: The Shaping of Disinformation at the Peripheries of Platform Ecosystems," *Social Media + Society* 10, no. 1 (January 1, 2024): 2, https://doi.org/10.1177/20563051231224721 < https://doi.org/10.1177/20563051231224721> .

16. Wendy Hui Kyong Chun, *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition* (MIT Press, 2021), 125, https://doi.org/10.7551/mitpress/14050.001.0001 < https://doi.org/10.7551/mitpress/14050.001.0001> .

17. Caitlin Petre, Brooke Duffy, and Emily Hund, "'Gaming the System': Platform Paternalism and the Politics of Algorithmic Visibility," *Social Media + Society* 5, no. 4 (2019): 2, https://doi.org/10.1177/2056305119879995 < https://doi.org/10.1177/2056305119879995> .

18. Elena Pilipets, Sofia P. Caldeira, and Ana Marta M. Flores, "'24/7 Horny&Training': Porn Bots, Authenticity, and Social Automation on Instagram," *Porn Studies* (July 18, 2024): https://doi.org/10.1080/23268743.2024.2362166 < https://doi.org/10.1080/23268743.2024.2362166> . See also Ysabel Gerrard and Helen Thornham, "Content Moderation: Social Media's Sexist Assemblages," *New Media & Society* 22, no. 7 (July 1, 2020): 1266–86, https://doi.org/10.1177/1461444820912540 < https://doi.org/10.1177/1461444820912540> ; Katrin Tiidenberg, "Sex, Power and Platform Governance," *Porn Studies* 8, no. 4 (September 27, 2021): 381-393, https://doi.org/10.1080/23268743.2021.1974312 < https://doi.org/10.1080/23268743.2021.1974312> .

19. Ariadna Matamoros-Fernández, Louisa Bartolo, and Betsy Alpert, "Acting Like a Bot as a Defiance of Platform Power: Examining YouTubers' Patterns of 'Inauthentic' Behaviour on Twitter During COVID-19," *New Media & Society* 26, no. 3 (February 26, 2024): 1290–1314, https://doi.org/10.1177/14614448231201648 < https://doi.org/10.1177/14614448231201648> .

20. Crystal Abidin, *Internet Celebrity: Understanding Fame Online* (Emerald Publishing Limited eBooks, 2018), https://doi.org/10.1108/9781787560765 < https://doi.org/10.1108/9781787560765> .

21. Douglas Guilbeault, "Growing Bot Security: An Ecological View of Bot Agency," *International Journal of Communication* 10 (October 12, 2016): 5003–21, https://ijoc.org/index.php/ijoc/article/view/6135 < https://ijoc.org/index.php/ijoc/article/view/6135> .

22. Petre, Duffy, and Hund, "Gaming the System," 8.

23. Kelley Cotter, "Playing the Visibility Game: How Digital Influencers and Algorithms Negotiate Influence on Instagram," *New Media & Society* 21, no. 4 (2018): 895–913, https://doi.org/10.1177/1461444818815684 < https://doi.org/10.1177/1461444818815684> .

24. Stefan Stieglitz, Florian Brachten, Björn Ross, and Anna-Katharina Jung, "Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts," *ACIS 2017 Proceedings* 89, (2017): https://aisel.aisnet.org/acis2017/89 < https://aisel.aisnet.org/acis2017/89> .

25. Petre, Duffy, and Hund, "Gaming the System."

26. Petre, Duffy, and Hund, "Gaming the System," 5.

27. Pilipets, Caldeira, and Flores, "24/7 Horny&Training."

28. Matamoros-Fernández, Bartolo, and Alpert, "Acting Like a Bot."

29. See for example Nathalie Schäfer, "1001 Followers in 20 Days: Framing the Playful Use of Fame Enhancing Bots on Instagram," *Transactions of the Digital Games Research Association* 6, no. 3 (August 13, 2024): 87–114, https://doi.org/10.26503/todigra.v6i3.2177 < https://doi.org/10.26503/todigra.v6i3.2177> . In this paper I analyze botting in detail. On the one hand, my analysis showed that Instagram and its users consider it cheating. Looking at different engagement practices, on the other hand, I question this perspective and call for further research to update the restrictive status of likes as an acknowledgment of the content, considering further intentions and meanings of liking and engaging in general.

30. Pilipets, Caldeira, and Flores, "24/7 Horny&Training."

31. Philip E. Agre, "Surveillance and Capture: Two Models of Privacy," *The Information Society* 10, no. 2 (April 1, 1994): 101–27, https://doi.org/10.1080/01972243.1994.9960162 < https://doi.org/10.1080/01972243.1994.9960162> .

32. Nurie Salim, "A Pornbot Stole My Identity on Instagram. It Took an Agonising Month to Get It Deleted,"*Guardian*, February 28, 2023, https://www.theguardian.com/lifeandstyle/2023/feb/28/a-pornbot-stole-my-identity-on-instagram-it-took-an-agonising-month-to-get-it-deleted < https://www.theguardian.com/lifeandstyle/2023/feb/28/a-pornbot-stole-my-identity-on-instagram-it-took-an-agonising-month-to-get-it-deleted> .

33. Pilipets, Caldeira, and Flores, "24/7 Horny&Training."

34. Anna Caterina Helm, "Was Steckt Hinter Den Sex-Bots Auf Instagram?" *Süddeutsche Zeitung Jetzt*, July 12, 2020, https://www.jetzt.de/digital/sex-bots-auf-instagram-das-steckt-dahinter < https://www.jetzt.de/digital/sex-bots-auf-instagram-das-steckt-dahinter> .

35. Paul Bouchaud, Raziye Buse Çetin, Natalia Stanusch, Salvatore Romano, and Marc Faddoul, "Pay-to-Play: Meta's Community (Double) Standards on Pornographic Ads," *AI Forensics,* January 8, 2025, https://aiforensics.org/work/meta-porn-ads < https://aiforensics.org/work/meta-porn-ads> .

36. Pilipets, Caldeira, and Flores, "24/7 Horny&Training." ↩

37. See Carolina Are, "An Autoethnography Of Automated Powerlessness: Lacking Platform Affordances in Instagram And Tiktok Account Deletions," *Media, Culture & Society* 45, no. 4 (December 12, 2022): 822–40, https://doi.org/10.1177/01634437221140531 < https://doi.org/10.1177/01634437221140531> ; Brooke Erin Duffy and Colten Meisner, "Platform Governance at the Margins: Social Media Creators' Experiences with Algorithmic (In) Visibility," *Media, Culture & Society* 45, no. 2 (2023): 285–304, https://doi.org/10.1177/01634437221111923 < https://doi.org/10.1177/01634437221111923> ; Stefanie Duguay, Jean Burgess, and Nicolas Suzor, "Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine," *Convergence* 26, no. 2 (2018): 237–52, https://doi.org/10.1177/1354856518781530 < https://doi.org/10.1177/1354856518781530> . ↩

38. "Entfernen nicht authentischer Aktivitäten auf Instagram." *Announcements* (blog),*Instagram,* November 19, 2018, https://about.instagram.com/de-de/blog/announcements/reducing-inauthentic-activity-on-instagram < https://about.instagram.com/de-de/blog/announcements/reducing-inauthentic-activity-on-instagram> . ↩

39. Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi, "Coordinated Behavior on Social Media in 2019 UK General Election," *Proceedings of the International AAAI Conference on Web and Social Media* 15 (May 22, 2021): 443–54, https://doi.org/10.1609/icwsm.v15i1.18074 < https://doi.org/10.1609/icwsm.v15i1.18074> ; Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu, "Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours," *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021): 1441–1451, https://doi.org/10.1145/3447548.3467391 < https://doi.org/10.1145/3447548.3467391> ; Timothy Graham, Sam Hames, and Elizabeth Alpert, "The Coordination Network Toolkit: A Framework for Detecting and Analysing Coordinated Behaviour on Social Media," *Journal of Computational Social Science* 7, no. 2 (May 11, 2024): 1139–60, https://doi.org/10.1007/s42001-024-00260-z < https://doi.org/10.1007/s42001-024-00260-z> . ↩

40. Christine L. Cook, Aashka Patel, and Donghee Yvette Wohn, "Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms," *Frontiers in Human Dynamics* 3 (2021): https://doi.org/10.3389/fhumd.2021.626409 < https://doi.org/10.3389/fhumd.2021.626409> . For further research into volunteer moderation and its challenges see: Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel, "Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities," *ACM International Conference on Interactive Media Experiences*, (2021): 61–72, https://doi.org/10.1145/3452918.3458796 < https://doi.org/10.1145/3452918.3458796> ; Beril Bulat, Hannah Wang, Stephen Fujimoto, and Seth Frey, "The Psychology of Volunteer Moderators: Tradeoffs between Participation, Belonging, and Norms in Online Community Governance," *New Media &*

Society 27, no. 11 (2024): 1–24, https://doi.org/10.1177/14614448241259028 < https://doi.org/10.1177/14614448241259028> ; Angela M. Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill, "Why Do Volunteer Content Moderators Quit? Burnout, Conflict, and Harmful Behaviors," *New Media & Society*, (2022), https://doi.org/10.1177/14614448221138529 < https://doi.org/10.1177/14614448221138529> .

41. Stefanie Duguay, Jean Burgess, and Nicolas Suzor, "Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine."

42. Kate Crawford and Tarleton Gillespie, "What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint," *New Media & Society* 18, no. 3 (2016): 410–28, https://doi.org/10.1177/1461444814543163 < https://doi.org/10.1177/1461444814543163> .

43. Crawford and Tarleton Gillespie, "What Is a Flag for?"

44. Duguay, Burgess, and Suzor, "Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine."

45. Sarah Myers West, "Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms," *New Media & Society* 20, no. 11 (2018): 4366–83, https://doi.org/10.1177/1461444818773059 < https://doi.org/10.1177/1461444818773059> .

46. Anatoliy Gruzd, Felipe Bonow Soares, and Philip Mai, "Trust and Safety on Social Media: Understanding the Impact of Anti-Social Behavior and Misinformation on Content Moderation and Platform Governance," *Social Media + Society* 9, no. 3 (2023), https://doi.org/10.1177/20563051231196878 < https://doi.org/10.1177/20563051231196878> .

47. Carolina Are, "Flagging as a Silencing Tool: Exploring the Relationship between De-platforming of Sex and Online Abuse on Instagram and TikTok," *New Media & Society*, (February 12, 2024), https://doi.org/10.1177/14614448241228544 < https://doi.org/10.1177/14614448241228544> .

48. *Kurz und heute*, "Neue Spam-Bot-Welle bei Instagram,"moderated by Till Haase, authored by Andreas Noll, aired January 31, 2019, on Deutschlandfunk Nova, https://www.deutschlandfunknova.de/beitrag/soziale-netzwerke-neue-spam-bot-welle-bei-instagram < https://www.deutschlandfunknova.de/beitrag/soziale-netzwerke-neue-spam-bot-welle-bei-instagram> .

49. Laestadius, "Instagram."

50. Tim Highfield and Tama Leaver, "A Methodology for Mapping Instagram Hashtags," *First Monday*, (December 26, 2014), https://doi.org/10.5210/fm.v20i1.5563 < https://doi.org/10.5210/fm.v20i1.5563> .

51. Geertz, "Thick Description."

52. Geertz, “Thick Description.” ↩

53. Geertz, “Thick Description.” ↩

54. Petre, Duffy, and Hund, “Gaming the System”; Schäfer, “1001 Followers in 20 Days.” ↩

55. Sofia Alexopoulou and Antonia Pavli, “‘Beneath This Mask There Is More Than Flesh, Beneath This Mask There Is an Idea’: Anonymous as the (Super)Heroes of the Internet?” *International Journal for the Semiotics of Law – Revue Internationale De Sémiotique Juridique* 34, no. 1 (2019): 237–64, https://doi.org/10.1007/s11196-019-09615-6 < https://doi.org/10.1007/s11196-019-09615-6> . ↩

56. Christian Fuchs, “The Anonymous Movement in the Context of Liberalism and Socialism,” *Interface: A Journal for and About Social Movements* 5, no. 2 (2013): 347. ↩

57. Blake Hallinan, Rebecca Scharlach, and Limor Shifman, “Beyond Neutrality: Conceptualizing Platform Values,” *Communication Theory* 32, no. 2 (2021): 201–22, https://doi.org/10.1093/ct/qtab008 < https://doi.org/10.1093/ct/qtab008> . ↩

58. Hallinan, Scharlach, and Shifman, “Beyond Neutrality.” ↩

59. See, for example, Alice E. Marwick, *Status Update: Celebrity, Publicity, and Branding in the Social Media Age* (Yale University Press, 2013). ↩

60. See Hallinan, Scharlach, and Shifman, “Beyond Neutrality: Conceptualizing Platform Values,” and Schäfer, “1001 Followers in 20 Days. Framing the Playful Use of Fame-Enhancing Bots on Instagram.” ↩

61. Indira Sen, Anupama Aggarwal, Shiven Mian, Siddharth Singh, Ponnurangam Kumaraguru, und Anwitaman Datta, “Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram,” *Proceedings of the 10th ACM Conference on Web Science,* (2018): 205–9, https://doi.org/10.1145/3201064.3201105 < https://doi.org/10.1145/3201064.3201105> . ↩

62. Alexopoulou and Pavli, “‘Beneath This Mask,” 237. ↩

63. Gilles Favarel-Garrigues, Samuel Tanner, and Daniel Trottier, “Introducing Digital Vigilantism,” *Global Crime* 21, no. 3–4 (2020), 189–95, https://doi.org/10.1080/17440572.2020.1750789 < https://doi.org/10.1080/17440572.2020.1750789> . ↩

64. José del Ama, “Das Schauspiel der Schande und der PR-Wert der Ehre. Anatomie eines Skandals,” *AugenBlick. Marburger Hefte zur Medienwissenschaft* (2011): 72, https://doi.org/10.25969/mediarep/2371 < https://doi.org/10.25969/mediarep/2371> . ↩

65. Gillespie, “Custodians of the Internet,” 207. ↩

66. Clive Emsley, *A Short History of Police and Policing* (Oxford University Press eBooks, 2021), https://doi.org/10.1093/oso/9780198844600.001.0001 <

https://doi.org/10.1093/oso/9780198844600.001.0001> .

67. Emsley, *A Short History of Police and Policing*.

68. Bundeszentrale Für Politische Bildung, "Polizei," bpb.de, October 12, 2021, https://www.bpb.de/kurz-knapp/lexika/politiklexikon/18048/polizei < https://www.bpb.de/kurz-knapp/lexika/politiklexikon/18048/polizei> .

69. Daniel Trottier, "Digital Vigilantism as Weaponisation of Visibility," *Philosophy & Technology* 30, no. 1 (2017): 55–72, https://doi.org/10.1007/s13347-016-0216-4 < https://doi.org/10.1007/s13347-016-0216-4> .

70. Are, "Flagging as a Silencing Tool: Exploring the Relationship between De-platforming of Sex and Online Abuse on Instagram and TikTok."

71. del Ama, "Das Schauspiel der Schande," 75.

72. Alaina Demopoulos, "Free the Nipple: Facebook and Instagram Told to Overhaul Ban on Bare Breasts," *Guardian*, January 18, 2023, https://www.theguardian.com/technology/2023/jan/17/free-the-nipple-meta-facebook-instagram < https://www.theguardian.com/technology/2023/jan/17/free-the-nipple-meta-facebook-instagram> .

73. Duguay, Burgess, and Suzor, "Queer Women's Experiences of Patchwork Platform Governance."

74. Alaina Demopoulos, "Free the Nipple."

75. Tom Divon, Carolina Are, and Pam Briggs, "Platform Gaslighting: A User-centric Insight Into Social Media Corporate Communications of Content Moderation," *Platforms & Society* (January 16, 2025): https://doi.org/10.1177/29768624241303109 < https://doi.org/10.1177/29768624241303109> .

Author Information

Nathalie Schäfer

Nathalie Schäfer is a PhD candidate in the research training group Media Anthropology at the Faculty of Media at Bauhaus-University Weimar, with expertise in the practice of "botting" with fame-enhancing bots on Instagram. Her research focuses on bots, cultural techniques and practices, the intersection and entanglement of human and automated (bot) cultures, social media, and platform

governance, as well as methodologies for researching digital media. Nathalie holds a BA in Art, Music, and Media from Philipps-University Marburg and the Université de Poitiers (2017), as well as an MA in Media Studies and a Master's degree in Cinéma et Audiovisuel from the Université Lumière Lyon II, the Bauhaus University Weimar, and the Universiteit Utrecht (2019). She is board member of the Association of Internet Researchers' (AoIR) executive committee and co-editor of *Szenen kritischer Relationalität"* (meson press 2024, with Bolwin et al.). Her web presence can be found at https://nathalieschaefer.de, and she is @nathalie@social.aoir on Mastodon.

View all of Nathalie Schäfer's articles.

Article details

Nathalie Schäfer, "Call the (Bot-)Police: User-Led Platform Governance of '(In)Authenticity' on Instagram," *Lateral* 14.2 (2025).

https://doi.org/10.25158/L14.2.13

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. Copyright is retained by authors.

Lateral is the peer-reviewed, open access journal of the Cultural Studies Association.

ISSN 2469-4053